Research Paper | Computer Science | E-ISSN No : 2454-9916 | Volume : 10 | Issue : 4 | April 2024

CREDIT CARD FRAUD DETECTION USING A STACKING ENSEMBLE APPROACH WITH LSTM AND RANDOM FOREST MACHINE LEARNING TECHNIQUES

Varun Chellapilla[1], Sravya Chikkam[2], Jayanth Sriram Melinati[3], Mr. M. Ezhilarasan[4]

[1, 2, 3] Puducherry Technological University, Puducherry

[4] Project Guide, Professor, PTU

ABSTRACT

Credit cards play an essential role in today's digital economy, and their usage has recently grown tremendously, accompanied by a corresponding increase in credit card fraud. Machine learning (ML) algorithms have been utilized for credit card fraud detection. However, the dynamic shopping patterns of credit card holders and the class imbalance problem have made it difficult for ML classifiers to achieve optimal performance.

This research project aims to develop a reliable credit card fraud detection system through a stacking ensemble method, integrating LSTM and Random Forest machine learning techniques. This approach aims to enhance fraud detection accuracy by leveraging the diverse strengths of both models. The system will undergo rigorous evaluation to ensure its efficiency in identifying fraudulent transactions while minimizing false positives. By combining the temporal sequencing capabilities of LSTM with the decision-making process of Random Forest, the proposed approach seeks to achieve heightened sensitivity to fraudulent patterns while maintaining computational efficiency. Ultimately, the objective is to bolster security measures and protect financial institutions and consumers from potential fraud risks.

KEYWORDS: Credit Card, Deep Learning, Ensemble Learning, Fraud Detection, Machine Learning, Neural Network.

INTRODUCTION

Information technology advancements have significantly impacted the financial sector, leading to the broad adoption of electronic commerce (e-commerce) platforms. Also, the recent outbreak of the novel coronavirus (COVID-19) pandemic has further shown the need for a more digital world and further expanded the e-commerce industry.

Artificial intelligence (AI) and machine learning applications in the financial sector can produce excellent results for companies, such as improved efficiency, reduced. operational cost, and enhanced customer satisfaction. Several ML-based systems have been developed to detect credit card fraud.

Meanwhile, building robust machine learning-based CCFD models has remained a challenge for some reasons. Firstly, conventional classifiers make predictions based on the transaction details only, such as amount, transaction country, and transaction type, ignoring the sequence of transactions that defines the clients' shopping behavior, which is useful in identifying appropriate fraud patterns. Secondly, credit card fraud datasets are highly imbalanced since genuine transactions significantly outnumber fraudulent transactions. Imbalance classification is a predictive modelling problem where there is an uneven distribution of samples across the classes. The class that makes up a large proportion of the dataset is called the majority class, while the class with a smaller proportion is called the minority class. Imbalance classification is a challenge because most ML algorithms were designed with the assumption of an even class distribution.

Therefore, to address these challenges, this project proposes a robust deep learning ensemble approach that combines long short-term memory (LSTM) with random forest as base models in a stacking ensemble framework, with a multilayer perceptron (MLP) as the meta-learner. This is done to introduce diversity in the base models using classifiers with different training methods.

Additionally, the hybrid synthetic minority oversampling technique and edited nearest neighbor (SMOTE-ENN) method are employed to balance the class distribution in the dataset.

MATERIALS AND METHODS

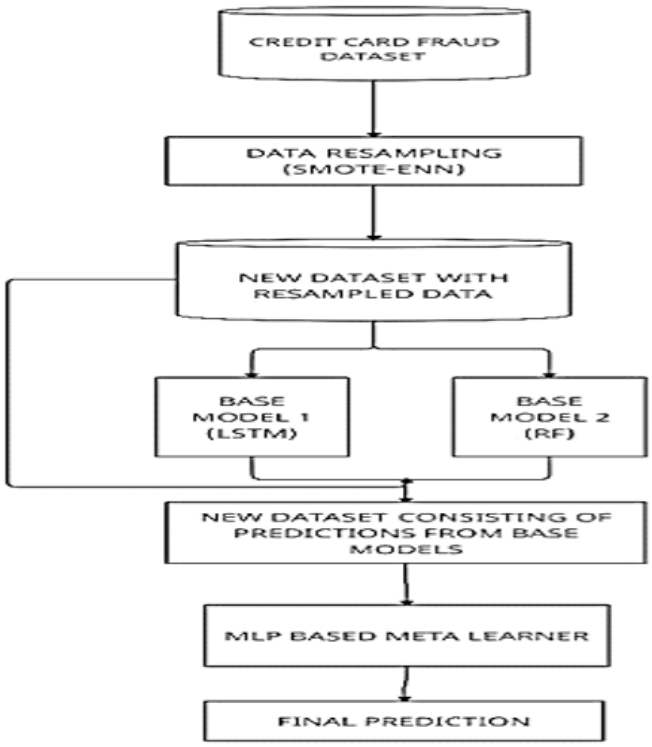


Fig.1: Structure of Proposed Model

Datasets:

Credit Card Dataset

This study uses a credit card dataset containing transactions performed by European cardholders in September 2013, which is publicly available. It contains 283,807 transactions, among which 492 transactions are labelled as fraudulent. The dataset is highly imbalanced, with only 0.172% labelled as fraudulent transactions. Most of the features were transformed to numerical variables using principal component analysis (PCA) because of confidentiality issues, and the names of the features were anonymized as V1, V2, V3, . . . , and V28, excluding the ''Time'' and ''Amount'' features. The ''Class'' feature is the target variable, and it has values 1 and 0, representing fraud and non-fraud transactions, respectively.

A. Multilayer Perceptron Multilayer perceptron is a type of feedforward neural network comprising three layers, including the input layer, hidden layer, and output layer. It is a powerful neural network with applications in several domains. In the MLP network, data flows from the input to the output layer. The hidden layer, placed between the input and output layers, is the core of the MLP, which processes the input information and transfers it to the output layer. The neurons are the processing elements in the MLP, and the neurons in each layer are connected to every neuron in the next layer. The input layer feeds the network with the input variables, and subsequent layers receive their inputs from the output of the previous layers. The MLP network is usually trained using the backpropagation algorithm, enabling the network to update its weights to minimize the output error. The mean squared error (MSE) is the commonly used error 30630 VOLUME 11, 2023 I. D. Mienye, Y. Sun: Deep Learning Ensemble with Data Resampling for Credit Card Fraud Detection function, and it is represented as:

$$E = 1/2\ n\Sigma\ i=1\ ||\ pi - ti\ ||\hat{}2 \quad (1)$$

where n is the number of data points, and pi and ti are the predicted output and target output for sample i, respectively. Meanwhile, this layer-to-layer transfer of information is achieved using activation functions, such as the sigmoid function $\sigma(k) = 1\ (1+e\ (-k))$, where e is Euler's number.

B. Long Short-Term Memory The LSTM network is a modified form of a recurrent neural network. Unlike conventional neural networks like MLP, RNNs are not limited to a unidirectional data flow. They can loop through several layers and temporarily memorize information that can be used later. Meanwhile, the simple RNN is susceptible to the vanishing gradient problem, and the LSTM and GRU were developed to solve the problem. The LSTM can learn long-term dependencies, making it suitable for classifying sequential data, such as credit card data. LSTM networks consist of a memory cell ct , with an input gate it , a forget gate ft , and an output gate ot . The three gates control how the data is processed and used. The following mathematical formulations represent the flow of information within the LSTM layers:

$$it = \sigma(Vixt + Wiht{-}1 + bi) \quad (2)$$
$$ft = \sigma(Vf\ xt + Wf\ h(t{-}1) + bf\ ) \quad (3)$$
$$c\tilde{}t = \tanh(Vcxt + Wch(t{-}1) + bc) \quad (4)$$
$$ct = ft \otimes c(t{-}1) + it \otimes \tilde{}ct \quad (5)$$
$$ot = \sigma(Voxt + Woh(t{-}1) + bo) \quad (6)$$
$$ht = ot \otimes \tanh(ct) \quad (7)$$

where V*, W*, and b* are learnable parameters, h* is the hidden state, where * is used in place of f, i, o, or c to represent the given gates and memory cell. Meanwhile, σ and tanh are the sigmoid and tanh activation functions and is the element-wise product.

C. Random Forest The Random Forest algorithm plays a pivotal role in the credit card fraud detection project due to its robustness, ensemble learning capabilities, and ability to handle complex datasets. Its robustness against overfitting ensures reliable performance even with intricate transaction data. By employing ensemble learning, Random Forest combines multiple decision trees, harnessing their collective power to capture diverse aspects of fraudulent patterns and improve overall detection accuracy. Additionally, its feature importance analysis aids in identifying the most influential features for fraud detection, facilitating a deeper understanding of fraudulent behavior. Random Forest's effectiveness in handling imbalanced data ensures equitable treatment of minority class instances, crucial for detecting rare fraudulent transactions amidst a sea of legitimate ones. Its scalability further enables efficient processing of large-scale credit card transaction datasets, making it an indispensable tool in the fight against financial fraud.

D. Ensemble Learing Ensemble learning is a machine learning approach that combines multiple algorithms to achieve better

classification performance than the individual base models [63] [64]. ML models usually have shortcomings, such as high bias, high variance, and low accuracy, and are not exempted from making errors. Therefore, rather than relying on one classifier, ensemble learning methods harness the strengths of two or more classifiers and often obtain higher accuracy than the individual base classifiers. Ensemble learning methods can be broadly grouped into bagging, boosting, and stacking. Stacking, which is the focus of this paper, involves using different base algorithms to build models, termed level-0 models, and a different algorithm called a meta-learner (or level-1 classifier) is trained to combine the predictions of the base models. The trained level-0 models are tested with out-of-sample instances, and the predicted class labels, combined with the actual labels, make up the dependent and independent variables in the new dataset employed for training the meta-classifier. Unlike bagging and boosting, which uses combination rules such as majority voting and weighted majority voting, the stacking ensemble model uses another ML algorithm (i.e., meta-learner) to aggregate the predictions from the level-0 models. In the literature, stacking-based ensembles have been applied in diverse fields.

MODULES DESCRIPTION:

Data Collection and Preprocessing: The dataset sourced from Kaggle's open-source platform, specifically from MLG - Machine Learning Group at ULB (Université Libre de Bruxelles), pertains to credit card fraud detection. This dataset is instrumental for research and development in the domain of machine learning and data science, particularly in the field of fraud detection and prevention within financial transactions. It comprises anonymized credit card transactions made by European cardholders over a period of two days, where each transaction is described by a set of numerical features representing various attributes such as time, amount, and principal components obtained through PCA transformation due to privacy concerns. The dataset is highly imbalanced, with a small fraction of transactions labeled as fraudulent, making it challenging for algorithms to detect fraudulent activities accurately. Therefore, it serves as an excellent benchmark for evaluating the performance of different fraud detection algorithms, facilitating the development of robust models capable of identifying fraudulent transactions while minimizing false positives, ultimately contributing to enhancing the security and integrity of financial systems.

Model Creation:

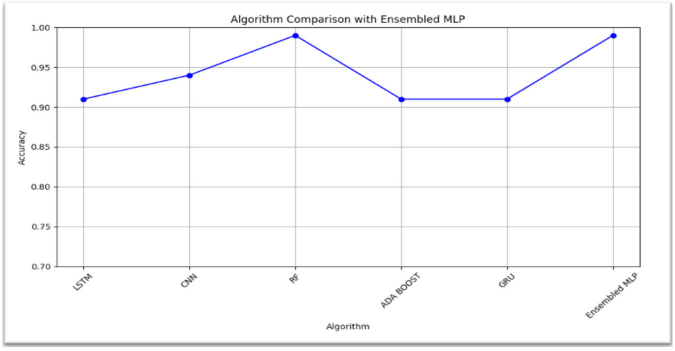
The modal training using LSTM and RF separately leverages the unique strengths of each approach to develop robust and accurate fraud detection models. While LSTM excels in capturing temporal dependencies and subtle patterns in sequential data, RF offers robustness and scalability, making it suitable for handling high-dimensional feature spaces and noisy datasets. By employing these methodologies separately, organizations can explore diverse modeling strategies and leverage the strengths of each approach to enhance the efficacy of their fraud detection systems.

Prediction and Evaluation:

A combined ensemble model integrating Random Forest (RF), Long Short-Term Memory (LSTM) and Multilayer Perceptron (MLP) techniques offers a potent solution for fraud detection. RF provides robustness and resilience to noise, LSTM captures temporal dependencies and subtle patterns in sequential data, while MLP captures intricate patterns and nonlinear relationships. By combining predictions from these models through techniques like averaging or stacking, the ensemble model achieves improved accuracy and generalization. This approach leverages the strengths of RF, LSTM and MLP, enhancing fraud detection performance and adaptability to evolving fraud patterns.

RESULTS AND DISCUSSION

This study develops a deep learning ensemble with data resampling for improved credit card fraud detection. The experimental results have been split into two, i.e., the classifiers' performance before and after data resampling. Meanwhile, the proposed stacking ensemble was achieved using the LSTM and Random Forest as base learners and an MLP neural network as the meta-learner.



CONCLUSION

In conclusion, the proposed approach for developing a Credit Card Fraud Detection (CCFD) system utilizing a stacking ensemble model showcases a comprehensive methodology aimed at enhancing fraud detection accuracy while ensuring user-friendliness through a seamless interface. By leveraging the strengths of LSTM and Random Forest as base learners and MLP as meta-learners within the ensemble framework, the system addresses the complexities of detecting fraudulent activities in credit card transactions effectively. The three-step process, starting with training base models using a robust 10-fold cross-validation technique to prevent data leakage, followed by the creation of a new dataset incorporating out-of-fold predictions and actual class labels, ensures the integrity and reliability of model training. Subsequently, the meta-classifier is trained using this dataset to combine predictions from base models, further enhancing the system's predictive capabilities. The development of a user-friendly interface adds practicality to the system, enabling easy utilization of the prediction system by stakeholders. This interface streamlines the process of accessing and interpreting fraud detection results, enhancing user experience and facilitating informed decision-making. Overall, the proposed approach not only focuses on improving fraud detection accuracy but also prioritizes usability and accessibility, making it a robust and practical solution for addressing the challenges associated with credit card fraud

detection in real-world scenarios.

FUTURE WORK:

In future work, enhancing the stacking ensemble model by incorporating additional diverse base learners and exploring advanced meta-learning techniques could further improve fraud detection accuracy. Additionally, extending the user-friendly interface to include real-time monitoring and interactive visualization features would enhance the system's usability and facilitate proactive fraud prevention measures. Moreover, investigating techniques to handle evolving fraud patterns and adaptively update the model would ensure the system's effectiveness in dynamic financial environments.

REFERENCES

1. I. D. Mienye and Y. Sun, "A Deep Learning Ensemble With Data Resampling for Credit Card Fraud Detection," in IEEE Access, vol. 11, pp. 30628-30638, 2023, doi: 10.1109/ACCESS.2023.3262020.
2. Dr Tina Elizabeth Mathew, "An Ensemble Machine Learning Model for Classification of Credit Card Fraudulent Transactions" in Journal of Theoretical and Applied Information Technology, Vol. 101 No. 9, pp. 3530-3546, 2023.
3. F. A. Ghaleb, F. Saeed, M. Al-Sarem, S. N. Qasem and T. Al-Hadhrami, "Ensemble Synthesized Minority Oversampling-Based Generative Adversarial Networks and Random Forest Algorithm for Credit Card Fraud Detection," in IEEE Access, vol. 11, pp. 89694-89710, 2023, doi: 10.1109/ACCESS.2023.3306621.
4. E. Esenogho, I. D. Mienye, T. G. Swart, K. Aruleba and G. Obaido, "A Neural Network Ensemble With Feature Engineering for Improved Credit Card Fraud Detection," in IEEE Access, vol. 10, pp. 16400-16407, 2022, doi: 10.1109/ACCESS.2022.3148298.
5. F. K. Alarfaj, I. Malik, H. U. Khan, N. Almusallam, M. Ramzan and M. Ahmed, "Credit Card Fraud Detection Using State-of-the-Art Machine Learning and Deep Learning Algorithms," in IEEE Access, vol. 10, pp. 39700-39715, 2022, doi: 10.1109/ACCESS.2022.3166891.
6. I.Benchaji, S.Douzi, B.Ouahidi, Jaafar Jaafari, "Enhanced credit card fraud detection based on attention mechanism and LSTM deep model." in Journal of Big Data 8, Article number: 151, 2021, doi:10.1186/s40537-021-00541-8
7. https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/iet-its.2020.0396
8. https://neptune.ai/blog/when-to-choose-catboost-over-xgboost-or-lightgbm
9. https://medium.com/@pushkarmandot/https-medium-com-pushkarmandot-what-is-lightgbm-how-to-implement-ithow-to-fine-tune-the-parameters-60347819b7fc.